Award Number: W81XWH-11-1-0549

TITLE: Enhancing the Breadth and Efficacy of Therapeutic Vaccines for Breast Cancer

PRINCIPAL INVESTIGATOR: Paul T. Spellman, PhD

CONTRACTING ORGANIZATION: Oregon Health Sciences University
Portland, OR 97201

REPORT DATE: June 2016

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved

OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
June 2016	Final	25 Sep 2011 - 24 Mar 2016

4. TITLE AND SUBTITLE

Enhancing the Breadth and Efficacy of Therapeutic Vaccines for Breast Cancer

5a. CONTRACT NUMBER

5b. GRANT NUMBER

W81XWH-11-1-0549

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Paul Spellman, PhD; Dmitri Rosanov, PhD; Kami Chiotti, MS

E-Mail: spellmap@ohsu.edu, rosanov@ohsu.edu, chiotti@ohsu.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Oregon Health Science University
Portland, OR 97201

8. PERFORMING ORGANIZATION REPORT

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The focus of the Spellman/Gray work group during the research period has been upon the generation of materials, tools, and data for the purpose of aiding and supporting the research and findings of the entire multi-team collaboration endeavoring to identify antigenic targets for breast cancer-infiltrating T cells. Our team has achieved a number of accomplishments. We have determined the likely specificity of immunogenic peptides for MHC alleles from a collection of MHC-I-bound epitopes eluted from the cell surface of twenty breast cancer cell lines. We developed computational pipelines to identify the sequence of the complete TCR heterodimer was applied to new samples sequenced by our colleagues in Denver. We also developed an RNAseq pipeline to identify likely neoantigens in breast cancer using public data.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					USAMRMC
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
U	U	U	UU	18	

Table of Contents

INTRODUCTION:

The OHSU Spellman/Gray work group is one of three collaborators funded by this Department of Defense Breast Cancer Multi-Team Award; the other two being comprised of the Lee work group from City of Hope (formerly of Stanford Medicine Cancer Institute) and the Slansky/Kappler work group from University of Colorado Denver/National Jewish Health. The major objective of this endeavor is to develop novel strategies aimed at the enhancement of the protective effects of anti-tumor T cells *in vivo* in a patient-specific manner based on the hypothesis that partially protective anti-tumor T cells exist within TDLNs in most breast cancer patients. This will be accomplished by identifying the antigens anti-tumor T cells target in different breast cancer subtypes, potentially including antigens preferentially expressed by breast cancer stem cells. We will identify both MHC-I- and MHC-II-restricted antigens driving both CD8 and CD4 anti-tumor T cells *in vivo*, as CD4 T cells are needed to optimally sustain vaccine-elicited CD8 T cells *in vivo* [1]. Identified antigens will be categorized as to breast cancer subtype-specificity or shared status amongst subtypes, with the intention a patient could be matched with an optimal set of vaccine antigens for her tumor. Another novel aspect of this project is the identification of altered peptides (mimotopes) that may more efficiently activate anti-tumor T cells than the natural tumor epitopes. A final objective is to identify small molecule anti-cancer agents that synergize with cytotoxic T lymphocytes (CTLs) to enhance immune-mediated killing. Collectively, this undertaking will produce a set of immunologically validated antigens and mimotopes for major breast cancer subtypes, and a set of agents that cooperate with immune killing. These can be used in combinations in a patient-specific manner to maximize clinical benefit while minimizing toxicity. The tools we develop will enhance the breadth and efficacy of existing and future approaches for immune therapy of breast cancer. We discuss here the Spellman/Gray group's specific efforts toward realizing the goals of this collaboration.

KEYWORDS:

Breast cancer, cytotoxic T lymphocytes, RNAseq, MiTCR, immune response, epitopes

OVERALL PROJECT SUMMARY:

Generation and initial analysis of T cell clones [Task 5]

Confirm tumor reactivity and HLA restriction of clones. Breast cancer therapy based on amplifying a patient's antitumor immune response depends on the availability of appropriate MHC class I-restricted, breast cancer-specific epitopes. To build a catalog of epitopes presented by breast cancer cells, we undertook systematic MHC class I immunoprecipitation followed by elution of MHC class I-loaded epitopes in breast cancer cell lines.

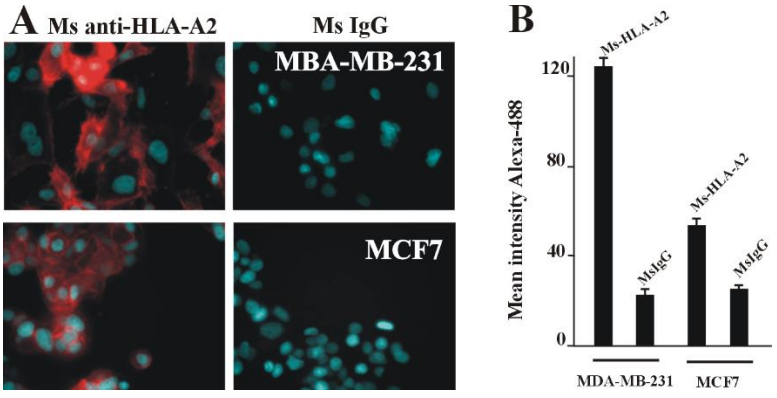

Figure 1. **A**, Immunostaining of HLA-A*02. MCF7 and MDA-MB-231 cells. **B**, Quantitative representation of HLA-A*02 expression in breast cancer cell lines.

First, we used immunohistochemistry to identify MHC class I positive breast cancer cell lines. In addition to MHC class I-positive breast cancer cells, we sought to identify HLA-A*02-positive cell lines because the HLA-A*02 allele occurs frequently in all ethnic groups. HLA-A*02 has been identified in 35% of African-Americans and in 50% of Caucasians [2]. Each cell line was stained with the MHC class I pan-antibody (clone W6/32) and HLA-A*02-specific antibody (clone BB7.2) followed by Alexa Fluor 488–conjugated donkey anti-mouse IgG. Control staining was performed with non-specific mouse IgG antibodies. Representative images of the stained MCF7 and MDA-MB-231 cells are shown in **Fig. 1A**. As expected, staining with control mouse IgG showed no signal. Typical quantitative data is shown in **Fig 1B**. The level of MHC class I expression in MDA-MB-231 cells was arbitrarily set to 100%, and the expression level of MHC class I in other lines was calculated as a percentage of MDA-MB-231 staining (**Table 1**).

Cell Line	Subtype	Relative to MDA-MB-231, %					HLA-A	HLA-B	HLA-C
		Protein		mRNA					
		HLA-A*02	MHC-I	HLA-A	HLA-B	HLA-C			
MDAMB231	Claudin-low	100	100	100 (287[a])	100 (146[a])	100 (109[a])	A*02:17/A*02:01	B*41:01/ B*40:02	C*17:01/ C*02:02
MCF7	Luminal	29	18	15	19	46	A*02:01/A*02:01	B*44:02/ B*18:01	C*05:01/ C*05:01
LY2	Luminal	17	31	13	28	61	A*02:01/A*02:01	B*44:02/ B*18:01	C*05:01/ C*05:01
HCC1500	Basal	39	25	93	139	215	A*02:02/A*02:02	B*15:80/ B*37:19	C*03:04/ C*04:01
SUM159PT	Claudin-low	26	21	98	68	67	A*24:02/A*02:01	B*51:01/ B*15:01	C*15:02/ C*03:03
BT549	Claudin-low	35	78	39	10	92	A*01:01/A*02:01	B*15:17/ B*56:01	C*07:01/ C*03:03
HCC1419	Luminal	2	7	15	36	126	A*24:02/A*02:01	B*46:01/ B*52:01	C*03:03/ C*01:02
CAMA-1	Luminal	4	15	14	44	86	A*02:01/A*32:01	B*40:02/ B*15:01	C*02:02/ C*03:03
MCF12A	Basal	21	26	50	43	149	A*66:01/A*02:01	B*41:02/ B*18:01/ B*35:08	C*17:01/ C*07:01
HCC1428	Luminal	5	6	38	59	319	A*01:01/A*02:01	B*07:02/ B*07:02	C*07:02/ C*07:02
UACC812	Luminal	25	29	55	89	245	A*68:01/A*02:05	B*15:03/ B*51:01	C*08:01/ C*12:03
HCC1395	Claudin-low	1	171	24	272	281	A*29:02/A*29:02	B*08:01/ B*45:01	C*07:01/ C*06:02
HCC1187	Basal	0	104	408	1708	1540	A*31:01/A*01:01	B*08:01/ B*40:01	C*07:01/ C*03:04
HCC1569	Basal	0	38	259	581	533	A*30:04/A*68:02	B*58:01/ B*53:01	C*04:01/ C*15:05
HCC70	Basal	0	54	406	2302	1394	A*30:02/A*03:01	B*78:01/ B*15:16	C*16:01/ C*16:01
MDAMB468	Basal	0	57	70	111	246	A*30:02/A*23:01	B*27:03/ B*53:01	C*02:02/ C*04:01
HCC1806	Basal	1	103	112	144	206	A*68:01/A*23:01	B*51:01/ B*15:03	C*02:02/ C*02:02
T47D	Luminal	0	46	35	587	392	A*33:01/A*33:01	B*14:02/ B*14:02	C*08:02/ C*08:02
SUM185PE	Luminal	0	100	0	58	184	A*36:01/A*36:01	B*40:01/ B*40:01	C*03:04/ C*03:04

Table 1. HLA phenotype and genotype in breast cancer cell lines. [a]Expected fragments per kilobase of transcript per million fragments sequenced (FPKM) [3].

MHC class I status in selected cell lines was also determined using RNA-seq analysis. The level of MHC class I mRNA expression is also presented in Table 1. The data showed that in most cases, MHC class I staining (protein expression) did not correlate with the level of MHC class I mRNA. In addition, MHC class I staining was below the limit of detection in MDA-MB-157, JIMT1, CAL-51, ZR75B, SKBR3, and BT474 cells despite the fact that MHC class I mRNA was present in these cells. We identified MHC class I alleles in the breast cancer cell lines by RNA-seq using seq2HLA [4] to map RNA-seq reads against a reference database of HLA alleles [5] (**Table 1**). There was general agreement between HLA-A*02 phenotype and genotype; however, some cell lines were phenotypically HLA-A*02-negative although HLA-A*02 mRNA was present.

To identify MHC class I-restricted epitopes expressed on the cell surface MHC molecules were immunoprecipitated and epitopes were eluted by acid treatment. Eluates from control IgG and anti-MHC class I samples were subjected to mass spectrometry (MS) analysis. IgG control eluate from HCC1187 cells contained few peptides (**Fig. 2A**). In contrast, anti-MHC class I eluate from HCC187 contained several hundred peptides (**Fig. 2B**). Immunoprecipitation and elution of peptides were repeated for all selected cell lines (**Table 1**) in independent experiments.

We used the PAW processing pipeline[6] to identify cell surface peptides and control peptide false discovery rates. The total numbers of eluted peptides and corresponding proteins are shown in **Table 2**. The data in Tables 1 and 2 suggest that high levels of cell surface MHC class I protein expression do not guarantee high levels of peptide loading. For example, despite three independent experiments, very few peptide epitopes were recovered from MDA-MB-231 cells, which expressed high levels of MHC class I. A total of 3,358 peptides derived from 3,070 proteins were eluted from MHC class I molecules immunoprecipitated from the surface of the 20 cell lines. After removal of duplicates, the total number of unique epitopes and corresponding proteins was 2,822 and 1,939, respectively.

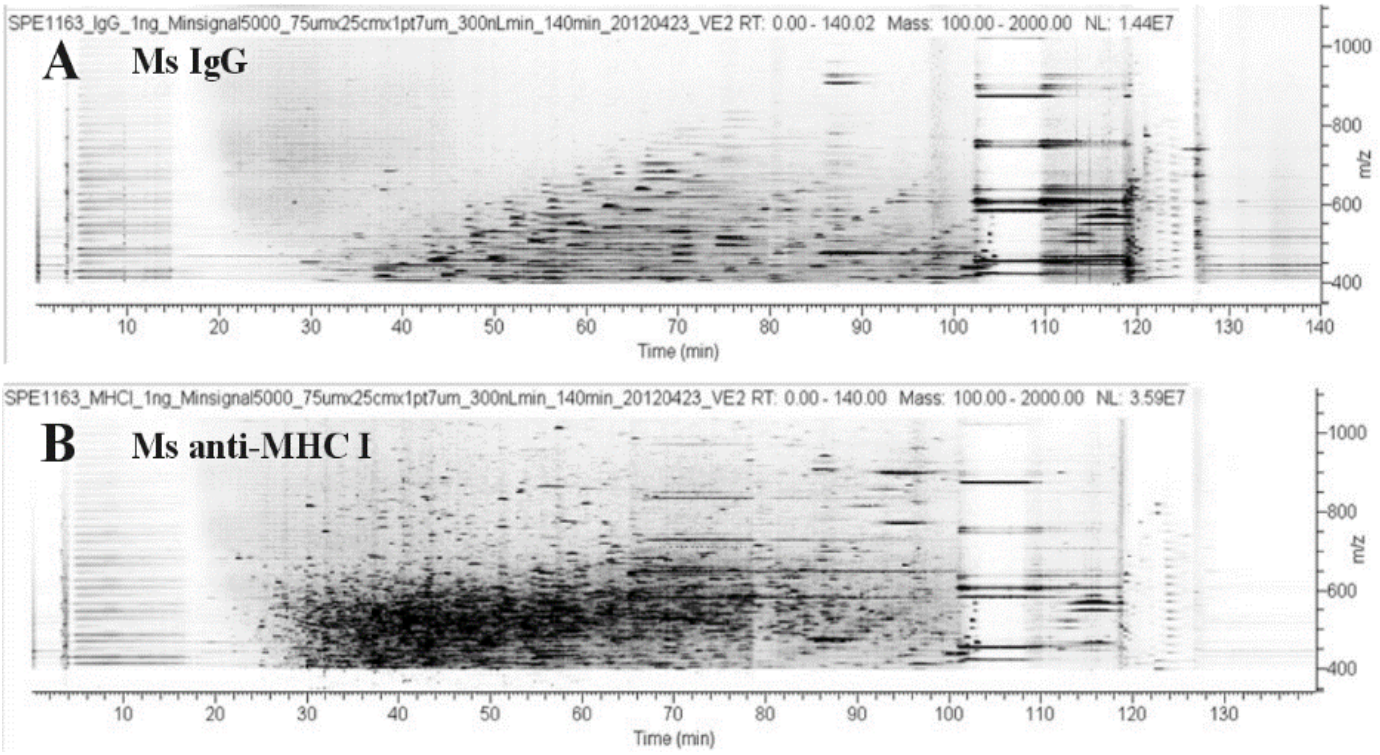

Figure 2. m/z vs time plots for HCC1187 with control and MHC class I antibodies. Mass spectrometry images of eluted peptides from nonspecific IgG (**A**) and anti-MHC class I (**B**) agarose incubated with HCC1187 cell lysate. Each dot on the chromatogram represents a single peptide. X axis: elution time; Y axis: mass to charge (m/z) ratio of the eluted peptides.

Next, we asked if the eluted epitopes contained mutated peptides, which are primary targets for T cell response. For this purpose we used specifically developed software, MS-Align+, which detects epitopes not only from normal proteins but also from mutated proteins and from proteins with new modifications[7]. We identified four epitopes derived from reverse translated genes tap1 (ATAPGLGGGPEPLGR) and ikbkap (EIISDPGVQGYSR) as well as from translations of a +1 frameshifted gene cp4 (AVASINSSEALR) and +2 frameshifted gene clipr1 (TAFESITSSDQR).

Analysis of our data showed that some peptides were most frequently presented on the cell surface of different breast cancer cell lines. As can be expected, peptides derived from proteins that are expressed at high levels in cancer and normal cells such as elongation factor 2 (EEF2), fructose-bisphosphate aldolase A (ALDOA), E3 ubiquitin-protein ligase RNF213 (RNF213), cytoplasmic dynein 1 heavy chain 1 (DYHC1), helicase with zinc

Sequence	Symbol	Gene	%CD137+	Cell Line
ALQEASEAYL	H3F3A	H3 Histone, Family 3A	4.5	MCF7
LLQEVEHQL	TRIM37	Tripartite Motif Containing 371	5.6	MCF7
HLFEKELAGQSR	LAD1	Ladinin 1	6.8	HCC1187
LLDVPTAAV	IFI30	Interferon, Gamma-Inducible Protein 301	8.0	MDAMB231, SUM159PT, MCF7, LY2
LLGPRLVLA	TMED10	Transmembrane Emp24-Like Trafficking Protein 10 (Yeast)	6.2	MCF7
AGAMAGVMGAYL	SLC25A35	Solute Carrier Family 25, Member 35	6.8	SUM159PT
AAAGSPVFL	SLC16A3	Solute Carrier Family 16, Member 3 (Monocarboxylic Acid Transporter 4)	4.3	MDAMB231
FTEAGLKELSEY	BZW1	Basic Leucine Zipper and W2 Domain-Containing Protein 1	4.5	HCC1187
AEIDAHLVAL	PSMA6	Proteasome Subunit Alpha Type-6	5.5	HCC1187
ILTDITKGV	EEF2	Eukaryotic Translation Elongation Factor 2	5.8	HCC1500, MCF12A, SUM159PT, LY2, MCF7
SAQGSDVSLTA	HLA-B	Major Histocompatibility Complex, Class I, B	8.0	SUM159PT, HCC70
No Peptide			2.9	

Table 2. Peptides that induce CD137 expression.

finger domain 2 (HELZ2), and eukaryotic translation initiation factor 3 subunit D (EIF3D) were also most frequently presented in the context of MHC class I.

We also worked on the characterization of epitopes eluted from breast cancer cells. To find breast cancer specific MHC I-loaded epitopes that have the ability to activate T cell response, we selected a subset of the eluted epitopes. First, we selected epitopes from genes that have alterations in at least 20% of invasive breast cancers using the cBioPortal. These alterations included copy number amplification, homozygous deletion, mRNA upregulation or downregulation, and mutation. Second, we used gene expression data from 708 breast tumor and 329 normal tissues from TCGA, EBI, and GEO as well as from 62 breast carcinoma cell lines and 6 non-transformed cell lines to identify epitopes from genes that have preferential expression in breast cancer samples over normal samples (at least 4 fold difference). We also selected epitopes and genes that were frequently identified by our MHC I immunoprecipitation and elution approach among different cell lines (at least 4 times).

Total number of epitopes that met the above-described criteria was 467. We were interested in HLA-A2-restricted epitopes, because HLA-A2 allele is the most frequent allele in the US and Caucasian population. Using specific software we determined HLA-A2 binding score for each peptide and selected peptides that have score at least 20. The highest score is 36. Approximately, 170 peptides were selected and synthesized for further analysis.

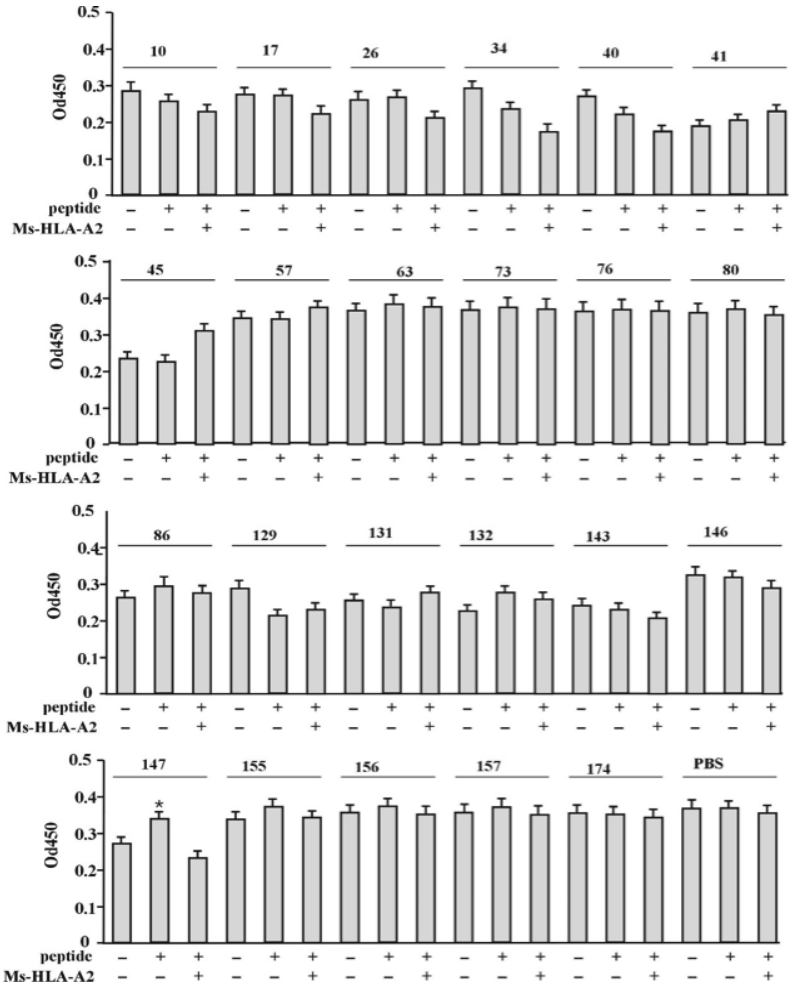

Figure 3. IFN-γ ELISA. Plate was coated with IFN-γ capturing antibodies (1 μg/ml, 100 μl/well) in PBS overnight at 4°C. C1R-A2 cells (10^5/well) were added to wells and pre-incubated with Ms-HLA-A2 (BB7.2) blocking antibodies (10 μg/ml) for 1 h. Next, C1R-A2 cells were loaded with peptide (10 μg/ml) and incubated 2 h followed by addition of T cells stimulated with corresponding peptide. C1R-A2/T cells mixes were incubated for 24 h and IFN-Y secretion was analyzed by IFN-γ-biotin antibodies and avidin-HRP. Signal was visualized by TMB-E and 1N HCL. Absorbance at 450 nm was measured using ELISA reader. *, $P = 0.05$.

To identify immunogenic peptides among selected peptides we used T cell activation protocol published in Blood (2007)[8] and J Immunol Methods (2006)[9]. Briefly, dendritic cells (DCs) were generated from HLA-A2-positive peripheral blood mononuclear cells (PBMCs) by a 90-min incubation at 37 °C in DC medium. Non-adherent cells and media were removed and replaced with 1 ml/well of fresh DC medium supplemented with 1000 IU/ml GM-CSF and 1000 IU/ml IL-4. After one day of incubation, DCs were matured using 10 ng/ml lipopolysaccharide (LPS) in the presence of peptide (10 μg/ml). Next day, peptide pulsed DCs were irradiated (32 Gy) and mixed with autologous CD8+ T cells and incubated for 7 days. On the day 4, IL-2 (50IU/ml) and IL-7 (5 ng/ml) were added to the medium. This stimulation was repeated 3 times.

Secondary stimulation was set up as described above, except that artificial antigen-presenting C1R-A2 were used. This secondary stimulation was repeated 3 times. After final stimulation, T cells were harvested and used in INF-γ ELISA with C1R-A2 cells loaded with peptide. **Figure 3** shows that peptide N 147 activates T cells as was evidenced by increased secretion of IFN-γ by T cells in the presence of peptide as compared to T cells without peptide or in the presence of HLA-A2 blocking antibodies.

The availability of comprehensive HLA-A, B, and C typing allowed us to predict the binding probability of each peptide to the HLA molecules present in the corresponding cells. For this purpose, we used a Consensus method[10] recommended by IEDB, which

consists of ANN[10],[11], SMM[12], Comblib[13], NetMHCpan[14], NetMHCcons[15], and PickPocket[16]. The predicted affinity of each peptide was expressed as an IC50 value, which was expressed as a percentile rank, generated by comparing the peptide IC50 against those of a set of random peptides from the SWISSPROT database. Lower rank values indicate higher predicted peptide affinities. Next, we calculated a percentile rank of binding to HLA-A*02 for each peptide, and plotted the number of peptides (counts) in relation to these percentile ranks (**Fig. 4**). We compared cell lines that were homozygous (MDA-MB-231, HCC1500, MCF7, and LY2), heterozygous (HCC1419, HCC1428, BT549, CAMA-1, MCF12A, UACC812, and SUM159PT), or null for HLA-A*02 based on the peptides that were derived from these cells in relation to percentile ranks (**Fig. 4**, *inset*). The results suggested that approximately 50% of the eluted peptides in HLA-A*02-positive homozygous cell lines were presented by HLA-A proteins.

To predict the distribution of HLA-A*02-specific peptides in breast cancer cell lines, we used a beta distribution

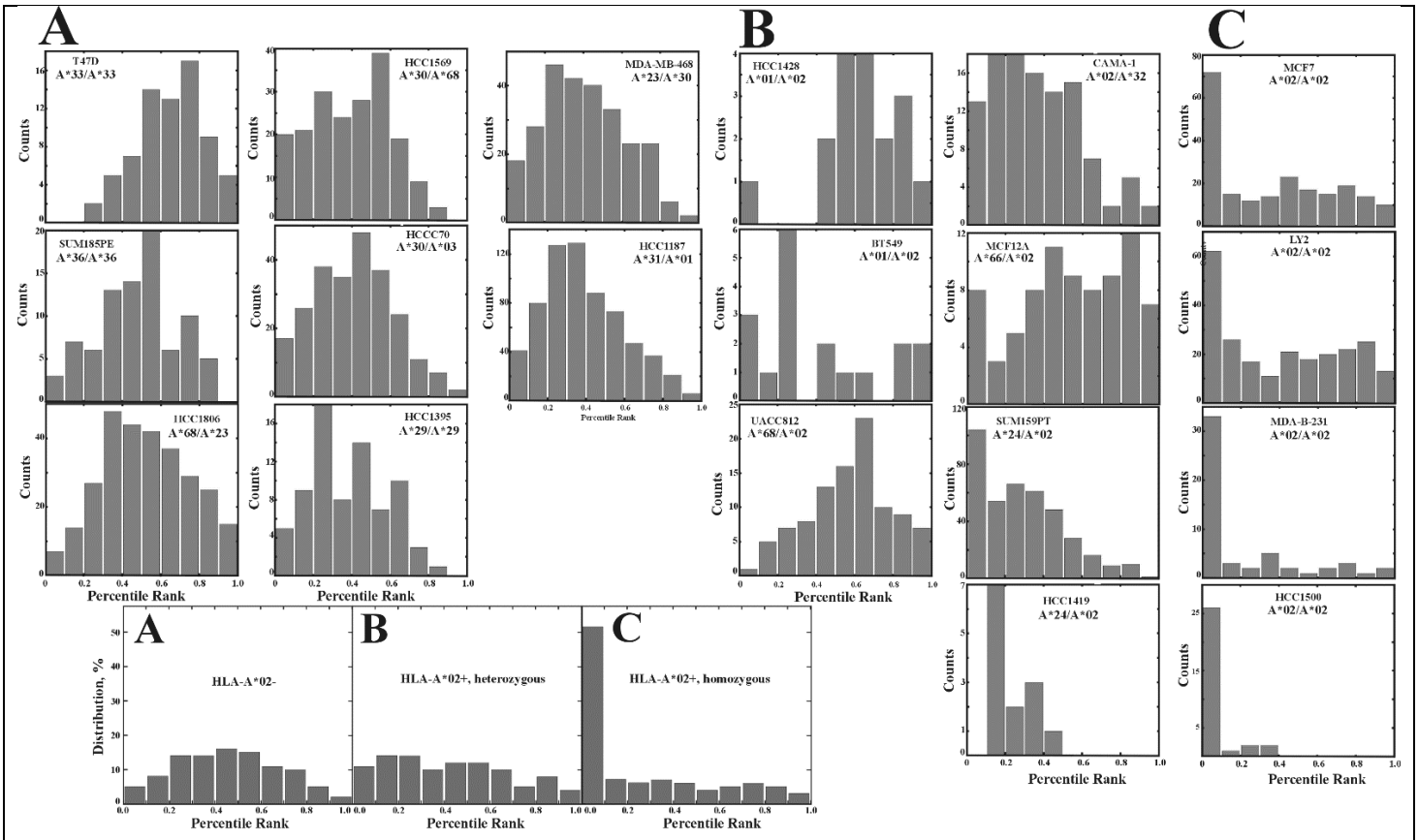

Figure 4. Distribution of peptides according to HLA-A*02 binding probability. Histograms showing distribution of peptides according to HLA-A*02 binding score in HLA-A*02-negative (**A**) and HLA-A*02-positive heterozygous (**B**) and homozygous (**C**) cell lines. Inset, histograms showing averaged relative distribution (percentage) of peptides in all HLA-A*02-negative (**A**) and HLA-A*02-positive heterozygous (**B**) and homozygous (**C**) cell lines.

mixture model[17], and fit the peptide distribution curve from the beta mixture model with Expectation Maximization (EM) implemented in Python. For HLA-A*02-negative cell lines, we modeled the data with one distribution to find the shape parameters for HLA-A*02-negative distribution. When modeling cell lines with an HLA-A*02-positive allele, we maximized the likelihood function to find the ratio between HLA-A*02-positive and HLA-A*02-negative alleles as well as the shape parameters for the HLA-A*02-positive distribution. This modeling process generated two beta distributions corresponding to HLA-A*02-positive and HLA-A*02-negative peptide distributions. In the follow up analysis, we determined the most likely proportion of HLA-A*02-positive peptides for each cell line from which we eluted at least 50 peptides (**Table 3**).

Cell line	FDR	HLA-A*02-	HLA-A*02+
MDA-MB-231	0.08	0.58	0.42
LY2	0.05	0.76	0.24
MCF7	0.06	0.71	0.29
CAMA-1	0.01	0.92	0.08
MCF12A	0.01	0.92	0.08
SUM159PT	0.13	0.79	0.21
UACC812	0.02	0.99	0.01
HCC1187	0.06	0.97	0.03
HCC1395	0.09	0.99	0.01
HCC1569	0.06	0.95	0.05
HCC1806	0.06	0.99	0.01
HCC70	0.09	0.96	0.04
MDA-MB-468	0.06	0.98	0.02
SUM185PE	0.02	0.98	0.02
T47D HER2+	0.01	1.00	0.00

Table 3. Ratio of HLA-A*02-positive and -negative peptides in breast cancer cell lines.

In addition, to determine if the eluted MHC class I-bound epitopes had been identified in previous studies, we searched the Immune Epitope Database (), which consists of 27,413 human unique peptide epitopes. We determined that of the 2,822 eluted unique epitopes, 803 epitopes have previously been shown to bind MHC class I proteins. Intriguingly, 9 of the peptides (GLLGTLVQL[18], ALSDHHIYL[18], SLFVSNHAY[19], SQFGGGSQY[19], NVIRDAVTY[19], VTAPRTLLL[20], VTAPRTVLL[20], ISDGPSKVTL[21], LLDVPTAAV[18], YGYDNVKEY[22]) in our data set were active in T cell activation assays (**Table 4**). Surprisingly, we found that these immunogenic epitopes were seldom related to breast cancer, and two of them were previously identified as targets of T cells in chronic lymphocytic leukemia (CLL) patients. This analysis demonstrated that both established cancer cell lines can be used as a source for the identification of tumor specific and immunogenic epitopes and immunogenic epitopes can be shared between different cancer types including leukemia and solid tumor cells.

In addition, in the course of these studies, we collected tumors and normal tissue from 25 HLA-A2+ positive patients followed by total RNA purification and RNAseq analysis (exome sequencing). RNAseq analysis will be used to identify overexpressed genes and mutations that can be presented in HLA-A2-loaded epitopes. RNAseq data will facilitate identification of immunogenic epitopes from cancer patients.

Peptide Sequence	Assay	Epitope source	MHC-I Restriction	Tumor Type	Number Identified
GLLGTLVQL	^{51}Cr release	Catenin β-1 [CTNB1] (400-408)	HLA-A*02:01	breast, ovarian, prostate	6
ALSDHHIYL	^{51}Cr release	Fructose-bisphosphate aldolase A [ALDOA] (216-224)	HLA-A*02:01	breast, ovarian, prostate, renal cell carcinoma	7
SLFVSNHAY	^{51}Cr release	Fructose-bisphosphate aldolase A [ALDOA] (356-364)	HLA-B*15:02, HLA-A*3003	in vitro, B cell line JS	1
SQFGGGSQY	^{51}Cr release	Eukaryotic translation initiation factor 3 subunit D [EIF3D] (61-69)	HLA-B*15:02; HLA-B*15:01;	in vitro; B lymphoblastoid cell line 721.221	2
NVIRDAVTY	^{51}Cr release	Histone H4 [H4] (65-73)	HLA-B*15:02	in vitro	0
VTAPRTLLL	^{51}Cr release	HLA class I histocompatibility antigen, B-37 alpha chain precursor [1B27] (3-11)	HLA-E	in vitro	0
VTAPRTVLL	^{51}Cr release	HLA class I histocompatibility antigen, B-50 alpha chain precursor [1B50] (3-11)	HLA-E; MHC-I(B) molecule Qa-1b	in vitro; H-2^{b} lymphoblasts	1
ISDGPSKVTL	Immunoblot detection of antibody/antigen binding	Coilin [COIL] (257-266)		in vitro	0
LLDVPTAAV	Positive MHC: epitope complex binding to TCR	Gamma-interferon-inducible lysosomal thiol reductase [GILT] (27-35)	HLA-A*02:01	in vitro; T2 cells; breast, ovarian, prostate; hepatocellular carcinoma (HCC) cell line SK-Hep-1 and solid tumors	4
YGYDNVKEY	ELISPOT IFN-γ release	CDCA7L protein [CDA7L] (422-430)	HLA-A*03, HLA-C*03, HLA-C*12	chronic lymphocytic leukemia (CLL)	7 +CLLs (23%)

Table 4. Immunogenic peptides that have been eluted from the cell surface of breast cancer cell lines.

TCR sequencing of each clone. Analysis was performed on sequencing data acquired from the Denver team's innovative emulsion rtPCR technique, yielding information about the identity of individual T cells through evaluation of paired TCR chains. This high-throughput data analysis was carried out using our modified version of miLaboratory's MiTCR TCR receptor repertoire analysis software, which was coined CompleteTCR.

The CompleteTCR pipeline was constructed to determine the repertoire diversity of T cell receptor clones from raw next generation sequencing data. CompleteTCR is built on the foundation of the MiTCR open source software package developed by MiLaboratory. MiTCR is a highly efficient and rapid approach to CDR3 extraction, clonotype assembly, and repertoire diversity estimation while accounting for sequencing and PCR errors as well as salvaging low-quality input reads. Currently, MiTCR is limited to analysis of either the α chain or the β chain (human or mouse) of the TCR heterodimer. CompleteTCR enhances the capabilities of MiTCR by allowing determination of TCR clone repertoire diversity of the matched αTCR-βTCR complex using the raw TCR sequence data of individual T cells generated by the Denver work group. This is accomplished via downstream manipulation of MiTCR outputs using R [23].

Since MiTCR assigns each input read a numeric identifier, it was necessary to make two modest changes to the MiTCR source code in order to produce the output required by CompleteTCR to match the α reads for each clonotype to their β mates. First, the standard MiTCR results file now includes a list of the numeric IDs for all reads belonging to each clonotype. Second, a temporary output file is created mapping the sequence identifier for each read in the input FASTQ file to its MiTCR-assigned numeric identifier. No changes were made to the algorithms MiTCR uses for CDR3 extraction, clonotype assembly, or error correction. The aforementioned R script first annotates the reads of each α clonotype with the appropriate sequence identifiers, repeating the process for the reads of each β clonotypes. The α and β reads are now paired according to their sequence identifier, and any read lacking a mate is removed from the dataset. Finally, the frequencies of αTCR-βTCR pairs, or clonotypes, are calculated.

This study resulted in the discovery of a TCR pair shared amongst 15 of 20 patient tumors and a number of other TCR pairs shared by 7 or more tumors. Since the TCR is examined ex vivo, data is not skewed by the effects of cell culture or cell death, lending confidence the shared TCR pairs are tumor-specific and hold potential as targets for immunotherapy. The false discovery rates of shared clonotypes was calculated, comparing the TCRs shared amongst the tumor PBMC samples to those shared amongst the normal/control PBMC samples. Of interest, we found the majority of statistically significant TCRs shared between the normal/control PBMC samples were absent from the entire set of tumor PBMC samples. Additional time was also spent ensuring any TCRs determined to be tumor-specific are, in fact, absent in the controls.

The OHSU team was also employed by the Denver team to perform additional computationally intensive analyses. For example, a custom script was designed to identify and quantify palindromic nucleic acid sequences residing in insertions at the VJ, VD, and DJ junctions of each clone identified (more than once) by CompleteTCR. These nucleotides, making up all or part of an insertion, inversely repeat the immediately adjacent germline DNA sequence.

Interpretations and implications of results from each study described in this section were performed by the Denver team and are currently in publication. We have made the CompleteTCR software package and source code freely available to the research community under a GNU license via GitHub [https://github.com/kamichiotti/CompleteTCR]. CompleteTCR requires Java version 1.7.0 [24] or higher and R version 3.1.0 [23] or higher with the *plyr()* package [25]. It is run from the command line via a shell wrapper script that requires an input manifest detailing locations of the α and β FASTQ files as well as their corresponding sample names. This approach facilitates high throughput data processing.

RNAseq analysis of tumor cells [Task 7]

RNAseq analysis to identify breast cancer-specific aberrant transcripts. Publically and privately available RNAseq datasets were used to conduct a systematic computational analysis identifying aberrant transcripts resulting in breast cancer antigens. The Spellman computational group developed an epitope prediction pipeline utilizing approximately 1000 breast cancer and normal tissue RNAseq samples available through The Cancer Genome Atlas (TCGA) [26], European Bioinformatics Institute (EBI) [27], and Gene Expression Omnibus (GEO) [28]. Over one-third of the RNAseq samples originated from normal adult tissues, predominantly

comprised of breast, lung, liver, brain, heart, kidney, and B-cells. A variety of other tissues are also represented, albeit in smaller sample numbers, to include bowel, skeletal muscle, lymph node, and ovary, amongst others. The entirety of the tumor dataset was obtained from the TCGA Data Portal. Of the better than 700 tumor samples, TCGA categorized approximately 460 samples into basal, Her2, and luminal subtypes using the PAM-50 subtype prediction method [29], enabling evaluation of subtype specificity of predicted transcripts. Only sequences generated on the Illumina Genome Analyzer II and Genome Analyzer IIx [30] platforms were included in the study to maintain as much uniformity as possible between datasets generated at different locations. As many of the sequences were single-end reads and read lengths varied from 50-150bp, the FASTQ files for all paired-end sequences were converted to single-end, and read lengths were trimmed as necessary to 50bp prior to submission to the analytical pipeline (Figure 5).
Initial mining of the RNAseq dataset was implemented via the Bowtie/Tophat/Cufflinks [3,31,32] packages (collectively referred to as the Tuxedo suite) to carry out sequence assembly and alignment to the human genome (hg19), prediction of novel isoforms, and quantitation of transcript structure. Using the Cuffmerge [3] feature of Cufflinks, the entire set of assemblies were merged such that identical transcripts across all samples were accounted for by a single identifier and its associated gene expression values.

Novel isoforms of a transcript can indicate alternative splicing events not yet characterized by the reference genome as well as aberrant structural variations due to mutation, both of which can result in neoantigens. Due to very low representation of the novel isoforms in some samples, it is likely the Tuxedo suite may not have detected, assembled, and subsequently determined the expression level for the new isoform in every sample. In order to force Tuxedo to look for and calculate the expression values of all isoforms in each sample, the subset of transcripts predicted to be novel assemblies were extracted from the Cuffmerge output and used to construct a new transcriptome index. The entire RNAseq dataset was rerun through the Tuxedo suite using this new index as the reference sequence. From here on, the collections of native and novel transcripts are kept separate from each other but run in parallel through the remainder of the pipeline.

For calculation of gene expression levels, we used the binary logarithm of the FPKM (fragments per kilobase of transcript per million mapped reads) values as calculated by Cufflinks. The FPKM values underwent full-quantile normalization utilizing the *betweenLaneNormalization* function of the EDASeq R/Bioconductor package [33]. This function accounts for distribution differences by matching the count distribution quantiles between samples, as described in [33] and [34]. Differential expression was evaluated across all normal and tumor samples by calculating the Median Split Silhouette (MSS) of each gene. MSS is a clustering algorithm measuring the average heterogeneity of possible clusters and determines whether the expression profile of a gene, across all normal and tumor samples, is best described by one or more clusters [35]. The advantage of MSS comes from its ability to identify biologically meaningful clusters where cluster size may be small. For our purposes, we limited the maximum number of potential clusters to three (*kmax*=3) in order to capture any distinction of gene expression between normal and tumor tissues as well as any

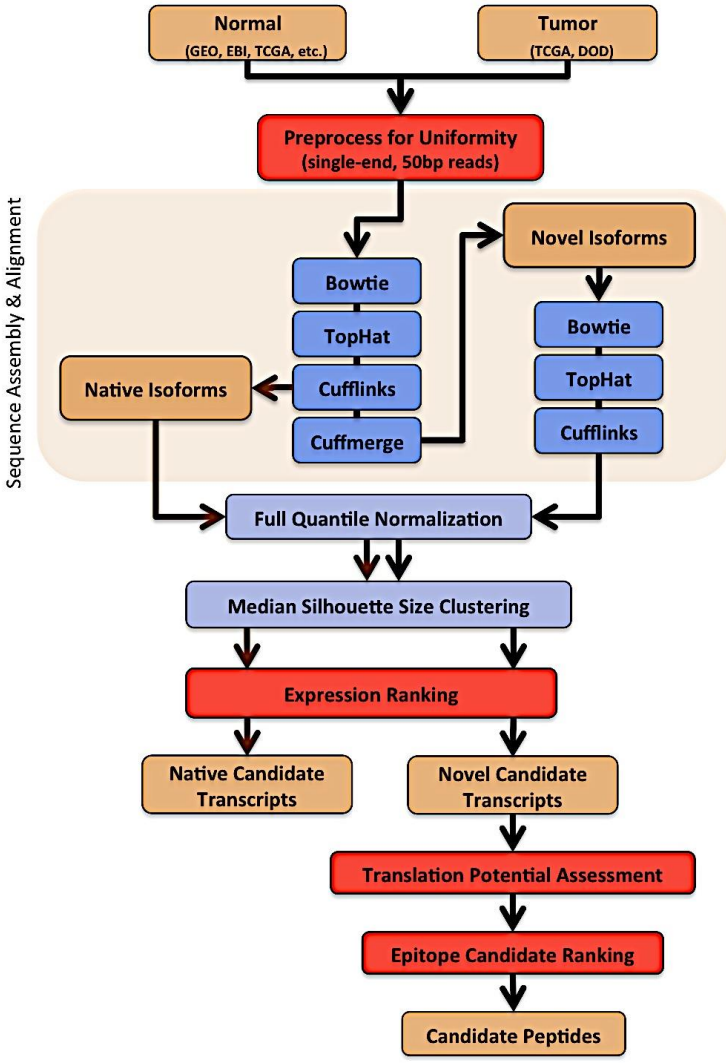

Figure 5. Pipeline for analysis of RNAseq data to identify native and neoantigen sequences.

bimodal expression within the tumor samples alone [36].

Originally, the strongest transcript candidates were selected by setting arbitrary cutoffs on percentage of tumor population and expression level represented in a cluster of interest. Transcripts failing to meet the set criteria were discarded. As we are also interested in where known immunogenic transcripts fall within the entire dataset, adjustments were made to the epitope discovery pipeline to establish 1) an enhanced method of ranking transcripts and epitopes with regard to expression specificity in tumor over normal tissues and 2) an automated workflow for discerning the unique portions of novel isoform sequences in large batches, rather than interrogating them individually.

To this end, a heuristic equation for *expression ranking* was established to calculate the rank of every candidate exhibiting a bimodal (high and low) or trimodal (high, mid, and low) expression profile across all samples. The equation is designed to highlight tumor-specific transcripts by placing higher weight on those where the high expression (H) cluster:

1) Is comprised predominantly of tumor samples as determined by the number of tumor samples (TS_H) present in the cluster population (CP_H):

   $$Tumor\ Fraction\ (TFx) = TS_H / CP_H$$

2) Represents a significant portion of the total tumor population (TP) represented by (TS_H):

   $$Tumor\ Population\ Fraction\ (TPFx) = TS_H / TP$$

3) Represents a minimal portion of the total normal population as determined by the complement of the total normal population (NP_H) represented by the number of normal samples (NS_H) present in CP_H.

   $$Complement\ Normal\ Population\ Fraction\ (CNPFx) = [1 - (NS_H / NP_H)]$$

4) Exhibits a significantly higher expression value than the low expression cluster as indicated by the difference between the unlogged medoid expression value of the high expression cluster (EV_H) and the of the low expression cluster (EV_L)

   $$Expression\ Difference\ (ED) = \alpha log(EV_H) - \alpha log\ (EV_L)$$

The priority ranking of each transcript candidate is then determined by:

$$Transcript\ Rank = TFx * TPFx * CNPFx * ED$$

The ranked novel assemblies now undergo *translation potential assessment* to elucidate those sequences possessing the best potential for translation into unique peptide constructs. The coding sequence of each transcript is translated in all three frames using the EMBOSS *transeq* tool [37,38], and the longest open reading frame (ORF) is selected. This sequence is aligned to the peptide sequences for all transcripts of the hg19 reference gene most closely related to the novel isoform using EMBL-EBI Clustal Omega [39] and EMBOSS *showalign* [37]. Any candidate lacking a start site shared by at least one of the reference sequences or possessing an ORF identical to any of the reference transcripts is removed from the dataset. Of the remaining candidates, the unique portion(s) of each transcript are aligned to the entire hg19 reference genome using BLAT [40], and any sequence(s) found to align elsewhere in the genome are also discarded. As longer candidate epitope sequences provide more opportunity for a true immunologic target, the remaining transcripts undergo *epitope candidate ranking* (**Fig. 1)** to take the length of the potential epitope into account.

$$Epitope\ Rank = Transcript\ Rank * Candidate\ Epitope\ Length$$

Twenty of the top-ranked known transcripts (**Table 5A)** and twenty of the top-ranked novel epitopes (**Table 5B**) are listed in **Table 5**. A number of the known transcripts provided in **Table 5A** are already known to be associated with breast cancer. The miR492 and miR622 micro-RNAs are found to have expression signatures correlated with specific breast cancer subtypes [41], and miR492 is particular is associated with supporting hepatic cancer progression through targeting of PTEN [42]. The cellular retinoic acid binding protein (CRABP2) is jointly regulated with estrogen receptor alpha and retinoic acid receptor alpha in human breast cancer cells [43]. The guanine nucleotide-binding protein subunit beta-2-like 1 (GNB2L1, or RACK1) has been reported as a predictor for pool clinical outcome in breast cancer patients and has potential to be an independent biomarker for diagnosis and prognosis of breast cancer. Upregulation of S100A11 is reported in a variety of metastatic cancers and is essential for the efficient repair of the plasma membrane and for the survival of highly motile

cancer cells [44], while overexpression of S100A14 modulates HER2 signaling in breast cancer [45]. Interferon alpha-inducible protein 6 (IFI6, or G1P3) promotes hyperplasia, tamoxifen resistance, and poor patient outcomes in breast cancer [46]. The estrogen-responsive anterior gradient 2 (AGR2) influences dissemination of metastatic breast cancer cells and may be useful as a marker in identification of circulating tumor and metastatic cells in sentinel lymph nodes. It is also a promising drug target and prognostic indicator [46].

A:

Gene	Low Expr (EV_L)	High Expr (EV_H)	Tumor Fxn (TF_X)	Tumor Popn Fxn (TPF_X)	Nrml Popn Fxn ($CNPF_X$)	Transcript Rank
TMSB10P1	9.71	12.36	0.85	0.39	0.13	1266.77
MIR492	0.00	11.45	0.74	0.99	0.66	698.78
RPL10	0.03	10.03	0.85	0.62	0.22	432.94
B2M	9.76	11.27	0.72	0.57	0.44	368.21
PABPC1	0.00	9.00	0.92	0.58	0.10	245.02
RPLP1	0.00	10.70	0.71	0.98	0.80	231.37
RPS24	0.00	9.26	0.80	0.72	0.36	226.93
CRABP2	1.80	8.45	0.97	0.69	0.04	219.88
GNB2L1	0.00	9.48	0.74	0.58	0.39	188.28
TFF1	0.15	8.98	0.93	0.42	0.06	185.09
RPL30	0.00	9.01	0.87	0.48	0.14	183.29
MYL6P1	8.04	10.22	0.71	0.41	0.33	179.36
RPL30	0.00	10.46	0.89	0.15	0.03	176.55
S100A11	5.79	9.25	0.76	0.98	0.59	168.61
MIR622	0.00	8.22	0.88	0.72	0.18	153.28
NPM1	0.00	8.70	0.80	0.70	0.34	152.20
S100A14	0.00	8.02	0.94	0.65	0.08	145.67
RPLP0	0.00	8.50	0.80	0.81	0.40	139.27
IFI6	6.06	8.69	0.92	0.47	0.08	136.30
AGR2	1.55	8.30	0.81	0.82	0.37	130.77

B:

Gene	Low Expr (EV_L)	High Expr (EV_H)	Tumor Fxn (TF_X)	Tumor Popn Fxn (TPF_X)	1-Nrml Popn Fxn ($CNPF_X$)	Transcript Rank	Epitope Length	Epitope Rank
TMSB10	11.25	12.74	0.86	0.82	0.26	2303.20	15	34548.03
KRT18	0.98	8.95	0.91	0.66	0.12	261.29	32	8361.19
ANXA2	7.07	10.22	0.77	0.96	0.58	325.58	12	3906.90
SEC61A1	0.00	6.70	0.80	0.57	0.28	34.06	100	3406.22
COL1A1	0.00	8.39	0.94	0.59	0.08	169.86	20	3397.28
MUC1	0.00	5.78	0.86	0.55	0.17	20.93	141	2951.29
SPINT2	2.61	7.20	0.77	1.00	0.59	44.53	61	2716.34
IGKV3-20	0.00	8.59	0.73	0.50	0.36	90.12	21	1892.46
SEC61A1	0.28	6.05	0.80	0.46	0.23	18.27	100	1827.15
TPD52	0.81	4.88	0.89	0.80	0.19	16.21	93	1507.84
GATA3	0.00	5.76	0.95	0.74	0.08	34.48	43	1482.68
TMED2	0.20	5.57	0.94	0.69	0.09	27.52	53	1458.58
HDLBP	4.41	6.51	0.81	0.46	0.21	20.70	67	1386.75
COL8A2	0.56	4.35	0.95	0.43	0.05	7.24	163	1179.65
HM13	2.00	5.85	0.94	0.57	0.08	26.59	44	1169.84
DDX23	0.01	4.74	0.73	0.55	0.41	6.07	176	1067.60
UGT2B11	0.00	7.40	0.87	0.08	0.03	11.66	86	1002.55
HNRNPM	0.00	5.43	0.85	0.56	0.19	16.13	62	1000.24
GTF2H5	0.00	8.27	0.77	0.69	0.41	96.30	10	963.00
LAMB2	1.88	5.14	0.60	0.40	0.53	3.52	243	854.16

Table 5. Twenty of the top ranked known (A) and novel (B) transcript candidates predicted by the epitope discovery pipeline in terms of 'transcript rank' for known transcripts and 'epitope rank' for predicted epitope sequence of novel isoforms.

The prevalence of breast cancer associated genes residing in high-ranking positions of this dataset lends significant support to the functionality of our pipeline as well as validity to the top-ranking epitope candidate results (**Table 5B**). In fact, even amongst the top-ranked epitope candidates shown in **Table 5B**, there are a number of cancer-related genes, including thymosin beta-10 (TMSB10, G-actin sequestration and breast

cancer cell motility) [48], keratin 18 (KRT18, tumor dedifferentiation and loss of estrogen and progesterone receptors) [48], and annexin A2 (ANXA2, invasion augmentation of multidrug-resistant breast tumor cells) [50].

As the predicted peptide sequence of a number of these candidates is the result of a single nucleotide change, sequencing of the aberrant region would be required for validation. Upon further inspection, the majority of the single nucleotide differences were determined to occur only at splice junctions, suggesting their prediction could be the result of erroneous alignment by Tophat, specifically alignment of 1-3 bases at the end of the read to the adjacent intron rather than to the correct position(s) at the beginning of the next exon. This issue has been addressed and corrected by more current spliced alignment programs, such as GSNAP, STAR, and Tophat2; however, the our dataset was originally aligned with Tophat in early 2013, overlapping the release date of Tophat2.

Before committing laboratory resources to candidate validation, we have chosen to reanalyze the TCGA breast tumors using an alternative alignment algorithm, Spliced Transcripts Alignment to Reference (STAR), to eliminate candidates predicted from poor splice junction alignment. This second pass of analysis involves the inclusion of a substantial number of additional RNA sequences to the original set of approximately 700 breast tumors (TCGA) and nearly 370 normal tissues (TCGA, GEO, EBI). We incorporated RNA sequence data for an additional >400 TCGA breast tumors, nearly 300 TCGA normal tissues, and >1700 GTEx normal tissues. This expansion of our normal tissue dataset provides an even more robust resource of normal transcript expression levels across all tissues against which tumors are compared for identification of uniquely expressed transcripts. With a dataset now more than double the original, along with the challenges encountered using an older alignment algorithm, it is imperative we reduce the computational resources required while improving alignment accuracy. To this end, implementation of STAR v2.4.2a for alignment to the most current version of the human reference genome (GRCh38) will have similar alignment accuracy and a runtime a fraction of that of Tophat2. The remainder of the pipeline remains unchanged. This analysis is still in progress. Candidates predicted via the epitope discovery pipeline, with support from both the Tophat and STAR algorithms, will be validated in the laboratory.

The current epitope candidate portfolio developed from this study is available upon request.

Identify small molecule agents enhancing tumor cell apoptosis and CTL killing [Task 12]
As outlined in Aim 4 of the proposal, clinical efficacy of T cell-based therapies will be enhanced in combination with agents promoting tumor cell apoptosis. Support for this idea recently has been published showing chemotherapy can synergize with CTL-mediated killing [51]; however, chemotherapeutic agents can also inhibit T cell function.

We are continuing our work in this area to identify drugs nontoxic to normal cells by developing T cell cytotoxicity assays using the peptides we previously characterized (see Task 5 above). We have determined the printing of Con A to specific spots on a slide allows attachment of PBMCs to these spots (**Fig. 6)**. We also

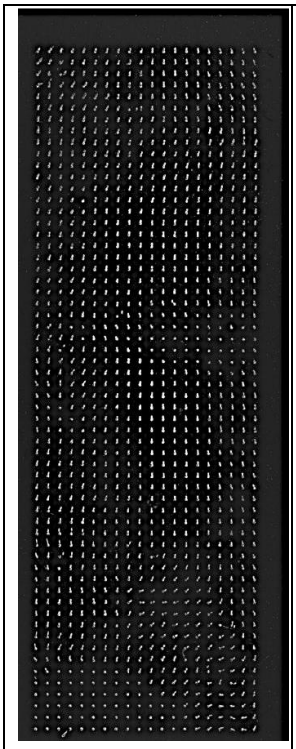

Figure 6. PBMC attachment to Con A spotted slide.

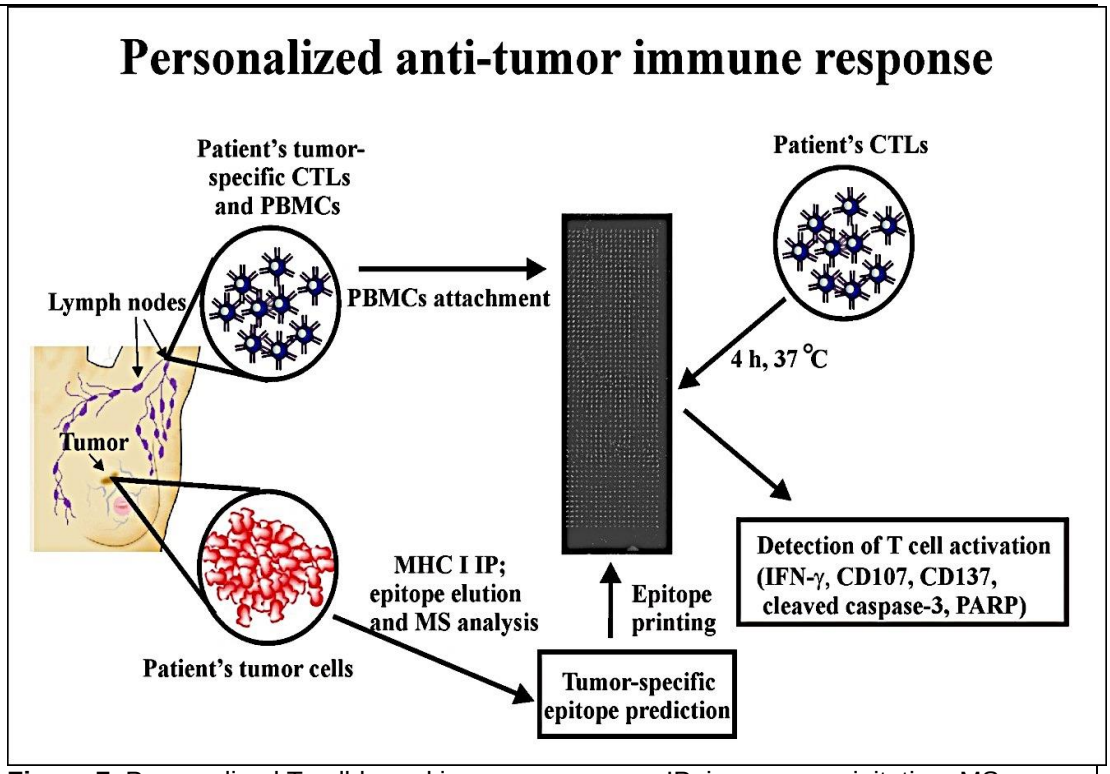

Figure 7. Personalized T cell-based immune response. IP, immunoprecipitation; MS, mass spectrometry

show attachment to Con A does not affect cell functionality, such as cell proliferation rate and ability of siRNA to reduce gene expression.

Finally, we designed a protocol to personalize T cell-based treatment (**Fig. 7)**. In this protocol, we will perform MHC I immunoprecipitation and epitope elution from patient tumor tissue, as we did with the breast carcinoma cell lines, followed by mass spectrometry analysis. Tumor-specific epitopes will be selected by gene expression analysis of the corresponding proteins. These epitopes, altogether with Con A, will be printed on a slide. Next, we will extract PBMCs and cytotoxic T lymphocytes (CTLs) from the same patient and allow the PBMCs to bind to the Con A spotted slide. Because the PBMCs are from the same patient, we do not need to know the type of MHC I alleles present in the patient. The slide, with attached PBMCs, will then be incubated with the patient's CTLs, and T cell activation will be detected using IFN-γ, CD137, CD107, and other T cell activation markers. We plan to test this protocol using tumors from breast cancer patients consented to the project by the City of Hope working group.

KEY RESEARCH ACCOMPLISHMENTS:

- Determined which of the 170 MHC I-loaded eluted epitopes identified in the previous funded year exhibited the ability to activate T cells. Eleven sequences were characterized as immunogenic, several of which were found in multiple cell lines.
- Modified open source MiTCR software to allow matching sequence reads from the alpha and beta chains of a single TCR followed by calculation of clonotype frequencies. Input to the program is raw sequence data from single TCRs generated by the Slansky team. The software is repackaged as CompleteClone.
- Modified the epitope discovery pipeline developed in the previous funding year for *in silico* prediction of breast cancer epitopes from RNAseq data. A more robust method for transcript and neoantigen candidate prioritization was instituted, and an automated approach for validating transcription potential of novel isoforms and isolation of potential neoantigen sequences was developed.
- Designing a protocol for personalization of T cell-based therapy through direct observation of tumor-derived T cell activation against epitopes eluted from the same patient.

CONCLUSION:

The focus of the Spellman/Gray work group over the past year has been upon the generation of materials, tools, and data for the purpose of aiding and supporting the research and findings of the entire multi-team collaboration endeavoring to identify antigenic targets for breast cancer-infiltrating T cells. We have identified a number of candidates in breast cancer tissues as well as breast cancer cell lines, utilizing a variety of analytical methods. The epitope discovery pipeline is proof of concept of *in silico* epitope discovery from RNAseq data. It aids in the definition of the protein-epitope relationship by enlarging the knowledge base of protein-encoding transcripts beyond the protein models existing in public databases and by restricting the analyses to only the expressed transcripts. The results produced by this pipeline along with the MHC-I-bound epitopes identified by mass spectrometry in breast cancer cell lines will be used to rank epitopes for further characterization and development as therapeutic targets.

PUBLICATIONS, ABSTRACTS, AND PRESENTATIONS:

A manuscript from the Denver team for which we are co-authors has been accepted at PNAS[52]. A manuscript describing CompleteClone is under development as is a manuscript describing the cancer cell line epitopes.

INVENTIONS, PATENTS, AND LICENSES:

No inventions, patents, or licenses to report.

REPORTABLE OUTCOMES:

NBCC/Artemis Project: We have developed a computational pipeline, coined CompleteClone, which analyzes raw TCR sequence data from single T cells, independently identifies the CDR3 sequence and VDJ alleles of the alpha and beta chains, matches the alpha and beta reads for individual TCR clonotypes, and calculates clonotype frequencies for the T cell clone. The software is currently used only with sequence data produced by the Slansky team following their single-cell emulsion RT-PCR technique; however, it can be packaged and shared for use with others for similar purposes.

OTHER ACHIEVEMENTS:

No other achievements to report.

REFERENCES:

1. Zajac, A. J., Murali-Krishna, K., Blattman, J. N. & Ahmed, R. Therapeutic vaccination against chronic viral infection: the importance of cooperation between CD4+ and CD8+ T cells. *Curr. Opin. Immunol.* **10,** 444–449 (1998).
2. Ellis, J. M. *et al.* Frequencies of HLA-A2 alleles in five U.S. population groups: Predominance of A∗02011 and identification of HLA-A∗0231. *Human Immunology* **61,** 334–340 (2000).
3. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28,** 511–515 (2010).
4. Boegel, S. *et al.* HLA typing from RNA-Seq sequence reads. *Genome Med* **4,** 102 (2013).
5. Boegel, S., Löwer, M., Bukur, T., Sahin, U. & Castle, J. C. A catalog of HLA type, HLA expression, and neo-epitope candidates in human cancer cell lines. *Oncoimmunology* **3,** e954893 (2014).
6. Wilmarth, P. A., Riviere, M. A. & David, L. L. Techniques for accurate protein identification in shotgun proteomic studies of human, mouse, bovine, and chicken lenses. *Journal of Ocular Biology, Diseases, and Informatics* **2,** 223–234 (2009).
7. Liu, X. *et al.* Protein Identification Using Top-Down Spectra. *Molecular & Cellular Proteomics* **11,** M111.008524–M111.008524 (2012).
8. Wolfl, M. *et al.* Activation-induced expression of CD137 permits detection, isolation, and expansion of the full repertoire of CD8+ T cells responding to antigen without requiring knowledge of epitope specificities. *Blood* **110,** 201–210 (2007).
9. Ho, W. Y., Nguyen, H. N., Wolfl, M., Kuball, J. & Greenberg, P. D. In vitro methods for generating CD8+ T-cell clones for immunotherapy from the naïve repertoire. *J. Immunol. Methods* **310,** 40–52 (2006).
10. Kim, Y. *et al.* Immune epitope database analysis resource. *Nucleic Acids Research* **40,** W525–W530 (2012).
11. Lundegaard, C. *et al.* NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res.* **36,** W509–512 (2008).

12. Peters, B. & Sette, A. Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method. *BMC Bioinformatics* **6,** 132 (2005).
13. Sidney, J. *et al.* Quantitative peptide binding motifs for 19 human and mouse MHC class I molecules derived using positional scanning combinatorial peptide libraries. *Immunome Research* **4,** 2 (2008).
14. Hoof, I. *et al.* NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* **61,** 1–13 (2009).
15. Karosiene, E., Lundegaard, C., Lund, O. & Nielsen, M. NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* **64,** 177–186 (2012).
16. Zhang, H., Lund, O. & Nielsen, M. The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics* **25,** 1293–1299 (2009).
17. Ji, Y., Wu, C., Liu, P., Wang, J. & Coombes, K. R. Applications of beta-mixture models in bioinformatics. *Bioinformatics* **21,** 2118–2122 (2005).
18. Barnea, E. *et al.* Analysis of endogenous peptides bound by soluble MHC class I molecules: a novel approach for identifying tumor-specific antigens. *Eur. J. Immunol.* **32,** 213–222 (2002).
19. Wei, C.-Y., Chung, W.-H., Huang, H.-W., Chen, Y.-T. & Hung, S.-I. Direct interaction between HLA-B and carbamazepine activates T cells in patients with Stevens-Johnson syndrome. *Journal of Allergy and Clinical Immunology* **129,** 1562–1569.e5 (2012).
20. García, P. *et al.* Human T cell receptor-mediated recognition of HLA-E. *Eur. J. Immunol.* **32,** 936–944 (2002).
21. Mahler, M., Mierau, R., Schlumberger, W. & Blüthner, M. A population of autoantibodies against a centromere-associated protein A major epitope motif cross-reacts with related cryptic epitopes on other nuclear autoantigens and on the Epstein-Barr nuclear antigen 1. *Journal of Molecular Medicine* **79,** 722–731 (2001).
22. Whitelegg, A. M. E. *et al.* Investigation of peptide involvement in T cell allorecognition using recombinant HLA class I multimers. *J. Immunol.* **175,** 1706–1714 (2005).
23. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2014).
24. Java SE - Downloads | Oracle Technology Network | Oracle. Available at: http://www.oracle.com/technetwork/java/javase/downloads/index.html. (Accessed: 6th May 2014)
25. Wickham, Hadley. The Split-Apply-Combine Strategy of Data Analysis. *J Stat Softw* **40,** 1–29 (2011).
26. Home - The Cancer Genome Atlas - Cancer Genome - TCGA. Available at: http://cancergenome.nih.gov/. (Accessed: 26th March 2013)
27. Kapushesky, M. *et al.* Gene Expression Atlas at the European Bioinformatics Institute. *Nucleic Acids Research* **38,** D690–D698 (2009).
28. Edgar, R. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **30,** 207–210 (2002).
29. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).
30. Illumina. (2013). Available at: www.illumina.com.
31. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
32. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
33. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-Content Normalization for RNA-Seq Data. *U.C. Berkeley Division of Biostatistics Working Paper Series* (2011).
34. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11,** 94 (2010).
35. Pollard, K. S. & Van Der Laan, M. J. A method to identify significant clusters in gene expression data. *UC Berkeley Division of Biostatistics Working Paper Series* (2002).
36. Bessarabova, M. *et al.* Bimodal gene expression patterns in breast cancer. *BMC Genomics* **11 Suppl 1,** S8 (2010).
37. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* **16,** 276–277 (2000).

38. Goujon, M. *et al.* A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res.* **38,** W695–699 (2010).
39. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7,** (2011).
40. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12,** 656–664 (2002).
41. Riaz, M. *et al.* miRNA expression profiling of 51 human breast cancer cell lines reveals subtype and driver mutation-specific miRNAs. *Breast Cancer Research* **15,** R33 (2013).
42. Jiang, J. *et al.* MicroRNA-492 expression promotes the progression of hepatic cancer by targeting PTEN. *Cancer Cell International* **14,** (2014).
43. Lu, M., Mira-y-Lopez, R., Nakajo, S., Nakaya, K. & Jing, Y. Expression of estrogen receptor alpha, retinoic acid receptor alpha and cellular retinoic acid binding protein II genes is coordinately regulated in human breast cancer cells. *Oncogene* **24,** 4362–4369 (2005).
44. Jaiswal, J. K. *et al.* S100A11 is required for efficient plasma membrane repair and survival of invasive cancer cells. *Nature Communications* **5,** (2014).
45. Xu, C. *et al.* S100A14, a Member of the EF-hand Calcium-binding Proteins, Is Overexpressed in Breast Cancer and Acts as a Modulator of HER2 Signaling. *Journal of Biological Chemistry* **289,** 827–837 (2014).
46. Cheriyath, V. *et al.* G1P3, an interferon- and estrogen-induced survival protein contributes to hyperplasia, tamoxifen resistance and poor outcomes in breast cancer. *Oncogene* **31,** 2222–2236 (2012).
47. Salmans, M. L., Zhao, F. & Andersen, B. The estrogen-regulated anterior gradient 2 (AGR2) protein in breast cancer: a potential drug target and biomarker. *Breast Cancer Res.* **15,** 204 (2013).
48. Mælan, A. E., Rasmussen, T. K. & Larsson, L.-I. Localization of thymosin β10 in breast cancer cells: relationship to actin cytoskeletal remodeling and cell motility. *Histochemistry and Cell Biology* **127,** 109–113 (2006).
49. Seon-Ah, H. *et al.* The prognostic potential of keratin 18 in breast cancer associated with tumor dedifferentiation, and the loss of estrogen and progesterone receptors. *Cancer Biomarkers* 219–231 (2011). doi:10.3233/CBM-2012-0250
50. Zhang, F. *et al.* Anxa2 plays a critical role in enhanced invasiveness of the multidrug resistant human breast cancer cells. *J. Proteome Res.* **8,** 5041–5047 (2009).
51. Ramakrishnan, R. *et al.* Chemotherapy enhances tumor cell susceptibility to CTL-mediated killing during cancer immunotherapy in mice. *J. Clin. Invest.* **120,** 1111–1124 (2010).
52. Munson, D. J. *et al.* Identification of shared TCR sequences from T cells in human breast cancer using emulsion RT-PCR. *Proc. Natl. Acad. Sci. U.S.A.* (2016). doi:10.1073/pnas.1606994113

APPENDICES:

No appendices to report.